ICS 35.040
CCS I 6540

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIAS 037—2025

智能算力云平台评估规范

Evaluation specification for cloud platform of intelligent computing capability

2025－11－12 发布　　2025－11－ 13 实施

上海市人工智能行业协会　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由上海市人工智能行业协会提出并归口。

本文件起草单位：上海智能算力科技有限公司、阿里云计算有限公司、上海人工智能创新中心、上海无问芯穹智能科技有限公司、上海仪电（集团）有限公司、上海市人工智能行业协会、上海埃迪希科技服务有限公司、上海埃迪西基础设施配套建设有限公司、中兴通讯股份有限公司、上海基流科技有限公司、上海壁仞科技股份有限公司、沐曦集成电路（上海）有限公司、上海天数智芯半导体有限公司、上海燧原科技股份有限公司、上海算丰信息有限公司、上海华东电信研究院、超聚变数字技术有限公司、上海科技网络通信有限公司

本文件主要起草人：孙跃、牛红星、王琳、辛帅、孙兆群、王翱、秦甘尧、刘俊豪、邹翔、王任杰、杨婷、王媞、罗高威、邱彭、陆欣、西羽、余健、张振华、曲振斌、裴芝林、钟普、吴保东、张骁立、赵旭、赵春昊、孟怀宇、于山山、杨毅、秦春华、左罗、冯晓磊、张国平、陶钰、夏宇、陈维、丁云帆、黄青青、彭莉、付轩、石加圣、邹翾、赵安璞、李超、梅敬青、华德宏、王思善、陈达亮、顾萌、罗捷、邓志辉、王超、陈香、钱涛、黄雷

本标准首次制定。

首期执行单位：上海无问芯穹智能科技有限公司、中兴通讯股份有限公司、上海基流科技有限公司

智能算力云平台评估规范

1 范围

本文件规定了智能算力云平台的评估框架、资源调度及管理功能评估、模型开发功能评估、模型应用功能评估、云平台性能评估和云平台安全评估的评估要求。

本文件适用于智能算力云平台的设计、开发、测试和运维，也可为智能算力平台的选型和评估提供参考依据。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867-2022 信息技术 人工智能 术语

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能集群 artificial intelligence cluster

遵循统一控制的，人工智能计算功能单元的集合。

[来源：GB/T 41867-2022，3.1.4]

3.2

智能算力云平台 cloud platform of intelligent computing capability

为智算集群提供资源调度及管理、模型开发、模型应用以及安全能力的云平台。

3.3

人工智能加速芯片 artificial intelligence accelerating chip

具备适配人工智能算法的运算微架构，能够完成人工智能应用运算处理的集成电路元件。

[来源：GB/T 41867-2022, 3.1.5]

3.4

预置大模型 preset large language model

平台预先集成、部署并优化好的生成式大模型服务，使用户可以直接调用或基于这些模型进行二次开发，避免用户从零开始训练或自行部署复杂的模型架构。

4 缩略语

下列缩略语适用于本文件。

API：应用编程接口（Application Programming Interface）

CPU：中央处理器（Central Processing Unit）

GPU：图形处理器（Graphic Processing Unit）

HTTP：超文本传输协议（Hyper Text Transfer Protocol）

HFU：硬件算力利用率（Hardware FLOPs Utilization）

MFU：模型算力利用率（Model FLOPs Utilization）

PEFT：参数高效微调（Parameter-Efficient Fine-Tuning）

QPS：每秒查询率（Queries Per Second）

RLHF：基于人类反馈的强化学习（Reinforcement Learning with Human Feedback）

SDK：软件开发工具包（Software Development Kit）

SFT：监督微调（Supervised Fine-Tuning）

SSH：安全外壳（Secure Shell）

5 评估框架

5.1 评估原则

5.1.1 客观性

被评估方应如实提供智能算力云平台评估要求的各项文件，确保文件的完整性、真实性和准确性。

评估方应客观、准确地对被评估方进行评估，对各项文件、相关资料进行评审、分析，真实准确地评估智能算力云平台的等级。

5.1.2 可追溯性

评估过程应有完整的文档记录。评估方应对支撑评估结果的文件进行归档、备案，确保相关结果可追溯。

5.1.3 保密性

评估方、被评估方应基于双方的保密要求，对评估过程中涉及的相关材料进行妥善保管和处理。

5.2 评估框架

智能算力云平台的评估框架见图1。智能算力云平台为人工智能集群提供资源调度及管理、模型开发和应用以及安全能力，平台的评估框架主要包括资源调度及管理功能评估、模型开发功能评估、模型应用功能评估、云平台性能评估和云平台安全评估。

智能算力云平台不同能力子域的测试方法示例见附录A。

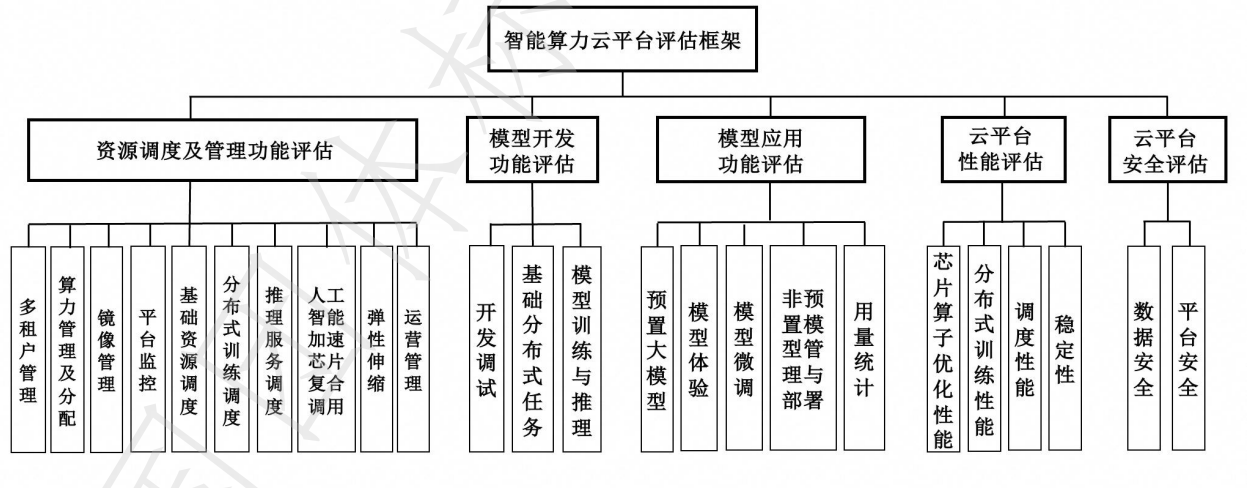

图1 智能算力云平台评估框架

5.3 评估等级和评估方法

智能算力云平台的评估等级分为基础级、提升级和引领级。基础级仅满足第6章～第10章必选的各能力子域要求（“应xx”条款）；提升级和引领级满足第6章～第10章必选的各能力子域要求（“应xx”条款），且满足部分或全部可选能力子域要求（“宜xx”条款），引领级满足的可选能力子域要求的数量应大于提升级。智能算力云平台的评估等级示例见附录B。

依据5.1的评估原则和第6章～第10章的评估要求，对各能力子域进行评分，计算出各能力子域的评分加和，由评估机构给出对应的评估等级。

5.4 评估流程

5.4.1 评估申请

被评估方按照评估要求，自愿向评估管理机构提交申请及支撑性材料，包括智能算力云平台等级自评报告及相关证明材料。智能算力云平台等级自评报告模版见附录C。

5.4.2 评估审查

评估管理机构委任评估专家组或委托第三方评估机构开展评估审查，包括：

a) 依据5.3的评估方法，对自评报告及相关证明材料进行技术审查；
b) 对需要现场确认的评估指标，进行实地检查和（或）测试验证；
c) 依据5.3的评估方法，对各能力子域进行符合性评定，计算出符合的条款数量和对应的评估等级，经过复核后，形成最终评估结果；
d) 宜明确评估结果的有效期。

5.4.3 评估报备

评估专家组或第三方评估机构向评估管理机构报备评估结果。

5.4.4 重新评估

在评估结果有效期到期、评估标准变动等条件下，可开展重新评估，重新评估流程应符合5.4.1、5.4.2和5.4.3的要求。

6 资源调度及管理功能评估

6.1 多租户管理

多租户管理的评估要求如下：

a) 应支持租户创建与管理、资源隔离、用户与权限管理等功能；
b) 应确保各租户之间的资源和数据隔离性；
c) 应支持灵活的资源分配和权限控制。

6.2 算力管理及分配

算力管理及分配的评估要求如下：

a) 应支持对服务器资源、异构算力资源、存储资源和网络资源的管理功能；
b) 应支持对资源的分配、资源监控；
c) 应支持对用户算力资源使用的全生命周期管理，包括从算力资源申请、创建、部署运行、计量到释放关闭的全生命周期管理；
d) 应支持按实际算力纳管与调度需求创建智算集群；
e) 应支持不少于5家人工智能加速芯片的资源管理；
f) 应支持高性能文件存储客户端集群部署及文件系统挂载；
g) 应支持基于集群和节点的计量能力，宜支持GPU、CPU、内存等资源的计量能力。
h) 宜支持大规模异构算力资源的纳管和调度能力。

6.3 镜像管理

镜像管理的评估要求如下：

a) 应支持对镜像上传、下载、版本管理以及在不同工作场景中的使用；
b) 应确保镜像管理的可操作性；
c) 应支持镜像加速能力，以便容器快速拉起；
d) 宜支持最小化拉取镜像中依赖的分层功能，提升实例启动速度。

6.4 平台监控

平台监控的评估要求如下：

a) 应支持租户监控功能，包括每个租户的各类人工智能加速芯片资源的可用数量和已使用数量等信息；

b） 应支持任务执行过程中的资源使用情况监控，包括任务执行过程中对CPU、人工智能加速芯片、内存等资源消耗监控功能和日志审计功能；
c） 应支持对智算集群的故障一键诊断；
d） 宜支持CPU、内存、存储、人工智能加速芯片、网络资源等维度的负载率监控；
e） 宜支持实时资源监控、任务资源监控、多维度监控以及告警功能，确保系统能够实时发现并处理资源消耗异常、性能瓶颈。

6.5 基础资源调度

基础资源调度的评估要求如下：
a） 应支持至少2种调度策略，如拓扑感知调度、优先级调度、故障感知调度等；
b） 应确保任务能够在多种调度策略下高效执行，并优化资源利用率；
c） 应具备人工智能加速芯片资源的拓扑感知能力，平台自动匹配最优的资源组合进行调度；
d） 宜支持利用统一的负载均衡或其他技术跨集群调用资源；

6.6 分布式训练调度

分布式训练调度的评估要求如下：
a） 应支持在分布式场景下多机多卡的资源调度和千卡规模调度；
b） 应支持当集群空闲节点未满足调度需求时任务可以正确被挂起；
c） 应支持弹性训练，支持配置弹性容错区间，在节点故障后调整节点数重新拉起任务；
d） 应支持当空闲节点满足调度需求时任务可以被正常调度；
e） 宜支持训练下网卡单网口故障下的高可用性，确保训练不中断；
f） 宜支持断点续训，支持弹性容错配置，配置后支持自动加载最近断点重新拉起训练任务；
g） 宜支持万卡规模调度。

6.7 推理服务调度

推理服务调度的评估要求如下：
a） 应支持在推理服务场景下的资源调度能力；
b） 宜支持调度扩展机制，保障用户资源配额的同时支持资源共享，提升集群推理资源利用率；
c） 宜支持多种卡的调度能力，例如1卡、2卡、4卡、8卡、16卡。

6.8 人工智能加速芯片复合调用

人工智能加速芯片复合调用的评估要求如下：
a） 应支持在人工智能加速芯片集群中选择CPU、内存资源规格；
b） 应支持人工智能加速芯片的共享复用、不同推理服务共享人工智能加速芯片的调度和不同实例被调度到同一张人工智能加速芯片的能力。

6.9 弹性伸缩

弹性伸缩的评估要求如下：
a） 应支持智算集群的扩缩容，按需增加或减少服务器数量；
b） 应支持租户资源的自动伸缩能力，如基于GPU指标进行自动扩缩容等；
c） 宜支持人工智能加速芯片的扩缩容，按需增加或减少人工智能加速芯片数量。

6.10 运营管理

运营管理的评估要求如下：
a） 宜提供向用户出售算力资源和解决方案的能力；
b） 宜支持为企业组织的决策系统提供信息支持，提供资源分析报告和合理操作建议；
c） 宜支持对各类云资源的统一运营管理，包括帐户管理、订单管理、帐单管理、经营分析等。

7 模型开发功能评估

7.1 开发调试

开发调试的评估要求如下：

a) 应支持利用开发机进行开发调试（包括网络服务终端、SSH远程登录和主流调试工具）；
b) 应支持在开发机内使用容器命令自定义环境；
c) 应支持在开发机内开放端口进行网络服务应用调试；
d) 应支持训练任务开发调试；
e) 应支持推理服务调试。

7.2 基础分布式任务

基础分布式任务的评估要求如下：

a) 应支持多机任务工具能力；
b) 应支持原生运行环境支持能力；
c) 应支持训练工具能力；
d) 应支持任务的复制、分配与管理能力；
e) 应支持任务的监控能力，能够查看任务运行过程的性能监控变化；
f) 应支持任务的日志能力，能够在网络服务终端查看任务输出的日志信息；
g) 应支持任务的算力负载节点的网络服务终端能力，方便调试；
h) 应支持使用高性能网络进行分布式任务；
i) 应支持使用高性能存储进行分布式任务；
j) 应支持使用分布式存储进行分布式任务；
k) 应支持分布式任务的生命周期管理，包括创建、暂停和继续。

7.3 模型训练与推理

模型训练与推理的评估要求如下：

a) 应支持对不同参数规模，混合专家模型和稠密模型的训练；
b) 应支持模型推理，确保平台在训练和多种推理框架下高效执行和稳定性；
c) 宜支持训练前的环境检测能力，确保训练任务启动在正常的计算资源上；
d) 宜支持训练容错能力，确保训练任务在发生异常时第一时间重新恢复训练，包括测试训练任务支持训前检测，测试过程中支持检查点自动备份策略配置，测试训练任务异常时是否进行错误检查并定位，测试是否重新调度算力负载节点并且从上一个检查点恢复训练，测试容错日志是否输出正常，测试是否支持最大10次的任务恢复次数。

8 模型应用功能评估

8.1 预置大模型

预置大模型的评估要求如下：

a) 应预置多种主流的大语言模型，覆盖模型介绍以及模型API；
b) 应支持通过提供模型类型、厂商、模型大小等标签维度对模型进行筛选；
c) 应涵盖多种主流语言HTTP、SDK的调用方式；
d) 应正确调用各类预置模型并正确返回相应结果；
e) 宜支持通过配置模型支持的多种模型采样参数，并获得相应的响应效果。

8.2 模型体验

模型体验的评估要求如下：

a) 前端交互模型能力应支持以对话、图像生成等方式体验；
b) 应支持对模型的参数配置进行修改；
c) 宜覆盖多模型在多种芯片上的效果和性能对比，确保平台提供便捷的免脚本的模型验证能力。

8.3 模型微调

模型微调的评估要求如下：

a) 应支持基础模型的选择、示例数据的查看、数据集的配置、训练参数的配置；
b) 应支持微调任务的状态查看、生命周期管理能力；
c) 应支持全参微调、PEFT、RLHF、蒸馏等方法；
d) 在输入的数据集符合平台要求的前提下，平台宜支持自动完成相应的模型微调任务。

8.4 非预置模型管理与部署

非预置模型管理与部署的评估要求如下：

a) 应支持用户自行上传、导入平台支持的大语言模型和多模态理解模型种类；
b) 应支持用户导入的模型的基本信息查看和生命周期管理；
c) 应支持基于微调服务生产的、或用户导入的符合平台要求的模型的部署服务；
d) 应支持部署基于量化任务生产的模型；
e) 应支持部署服务的生命周期管理和服务调用能力；
f) 应支持用户导入的生图模型被用于用户导入的图像工作流使用；
g) 宜支持用户部署定制模型，并可用与调用预置模型API类似的方式，调用部署的模型服务。

8.5 用量统计

用量统计的评估要求如下：

a) 应统计不同时间段大语言模型、多模态模型等调用的模型范围和调用产生的token（词元）数量；
b) 应统计不同时间段用户发起的模型微调任务数量、训练使用的token数量以及训练时长；
c) 应统计不同时间段用户调用的工作流的数量、任务次数、状态、运行时长等关键信息；
d) 宜支持多维度的功能模块用量统计能力。

9 云平台性能评估

9.1 芯片算子优化性能评估

芯片算子优化性能评估要求如下：

a) 平台应支持不同型号芯片算子，并确保GPU运算性能、正确性、稳定性；
b) 平台宜支持网络通信库的优化；
c) 平台宜支持对通信库的通信优化。

9.2 分布式训练性能评估

分布式训练性能评估要求如下：

a) 应支持同构同集群训练，在不同训练框架下测试多种模态模型（多模态理解模型/大语言模型）在同构同集群环境下不同型号芯片上训练的硬件利用率HFU及吞吐率；
b) 应支持在训练过程中进行训练过程观测，包含损失值、梯度值等；
c) 应支持训练性能分析，评估特定模型在分布式训练过程中产生的性能抖动，性能指标包括训练任务期间各周期各阶段用时分布与走势；
d) 应支持训练任务对不同型号芯片的物理拓扑感知；
e) 应支持分布式训练任务在网元、网际链路、GPU卡等维度的性能可视化分析。

9.3 调度性能评估

调度性能评估要求如下：

a) 应支持单集群内多个不同算力子任务并行进行，评估集群性能拓展性；
b) 应支持调度性能评估，测试的调度性能指标包括GPU利用率、多卡集群加速比、单周期吞吐速率、单周期各阶段用时占比。

9.4 稳定性评估

9.4.1 算力稳定性

稳定性的评估要求如下：

a) 云平台应确保在容器、GPU、操作系统、网络等资源层面具备诊断能力；
b) 云平台应确保在硬件故障情况下的响应能力和恢复速度；
c) 云平台应确保在实际工作负载下的可靠性和鲁棒性；
d) 云平台宜确保容错功能的自动故障检测和重新调度。

9.4.2 平台稳定性

平台稳定性的评估要求如下：

a) 云平台应确保容器化Linux实例的各个生命周期阶段的稳定性，包括启动、运行和管理；
b) 云平台应确保资源分配的准确性和效率（通过测试人工智能加速芯片的挂载和使用情况）；
c) 云平台应确保训练任务的资源可用性链路的拓扑可视化能力；
d) 云平台应确保在大规模训练场景（千卡以上）下节点的资源监控有效性；
e) 云平台宜确保持续负载下的稳定性，包括资源管理、负载均衡和系统响应等方面。

10 云平台安全评估

10.1 数据安全

数据安全的评估要求如下：

a) 云平台应提供数据访问控制和权限管理，提供数据在存储和传输过程中的数据加密；
b) 云平台应具备数据备份和恢复机制，确保在发生意外时能快速恢复重要数据；
c) 云平台应支持数据资源接入，数据识别、数据脱敏、数据审计、敏感数据保护等功能。

10.2 平台安全

平台安全的评估要求如下：

a) 云平台应具备安全防护体系，包括物理设施安全、操作系统安全、虚拟化安全、应用安全、运维安全、账号安全等；
b) 云平台应提供公网访问接口的安全防护机制；
c) 云平台应提供平台的用户认证和授权机制；
d) 云平台应提供容器的隔离性，避免存在容器逃逸风险；
e) 云平台应提供网络安全策略和防火墙配置功能，确保有效防御网络攻击，识别和修复潜在的安全漏洞；
f) 云平台应提供密钥管理能力，并支持用户自主创建、管理、使用密钥；
g) 云平台宜根据组织架构和资源集进行权限、资产的管理控制，支持资产导入、资产管理、资产识别以及风险分析处置。

附录 A
（资料性）

智能算力云平台测试方法示例

A.1 资源调度及管理平台功能评估

A.1.1 多租户管理

能力子域	测试目标	测试内容	预期结果
租户创建与管理	验证平台的租户创建、修改和删除功能，确保租户信息管理的准确性和完整性。	租户创建：测试管理员创建新租户的功能，检查租户名称输入、管理员分配、资源配额设置等操作。	平台能够成功创建租户，租户能够登录云平台。
		租户修改：测试租户信息（如描述、资源配额）的修改功能。	所有操作均能在系统中正确反映，不影响其他租户的正常运行。
资源隔离与配额管理	验证系统在多租户环境下的资源隔离性和配额管理功能，确保各租户之间的资源分配合理且互不干扰。	资源隔离：测试各租户之间的计算资源、存储资源和网络资源的隔离性。	各租户资源使用相互独立，无资源冲突或干扰。
		配额管理：测试 CPU、GPU、内存、存储配额的有效性，验证配额限制的提示和应对机制。	配额管理有效控制资源使用，配额超限时系统能够及时发出警告。
用户与权限管理	验证用户与权限管理功能，确保不同用户角色的操作安全性和资源使用权限。	用户角色管理：测试多级角色管理功能，验证不同角色的权限分配与操作范围。	平台正确管理用户角色和权限，不同角色的用户只能访问和操作其权限范围内的资源。
		用户管理：测试用户的添加、删除和角色变更功能。	验证用户的增删改查能力
		访问密钥管理：验证用户 SSH 公钥的配置和管理功能，确保远程访问的安全性。	SSH 公钥配置与管理功能安全可靠。

A.1.2 平台监控

测试项目	测试目标	测试内容	预期结果
实时资源监控	验证平台的实时资源监控功能，确保系统能够及时反馈资源使用情况，快速发现异常。	GPU 监控：测试对 GPU 资源的实时监控功能，确保系统能够准确反馈每个GPU的使用情况和性能指标。	系统能够实时监控每个 GPU 的使用情况，准确反馈 GPU 负载、温度、使用率等性能指标。
		存储监控：测试对存储资源的实时监控功能，确保系统能够监控存储空间的使用情况及输入/输出性能。	系统能够实时监控存储空间的使用情况和 I/O 性能，及时发现并报告潜在的存储瓶颈或异常。
		网络监控：测试对网络资源的实时监控功能，验证系统对网络负载、带宽使用和延迟的监控能力。	系统能够实时监控网络资源的使用情况，准确反馈网络负载、带宽使用和延迟等性能指标。
任务资源监控	验证平台对任务执行过程中的资源使用情况进行监控，确保任务执行的高效性和稳定性。	资源消耗监控：测试任务执行过程中对 CPU、GPU、内存等资源的消耗监控功能，确保用户能够实时了解任务的资源使用情况。	系统能够实时监控任务执行过程中的资源消耗情况，用户能够查看到每个任务的资源使用细节。
		吞吐量监控：测试任务执行中的吞吐量监控功能，确保系统能够反馈任务处理效率及其变化情况。	系统能够实时监控任务的吞吐量，准确反馈任务处理效率及其变化趋势。
		日志监控：测试对任务执行日志的实时监控功能，确保用户能够及时查看和分析任务执	用户能够实时查看任务执行过程中的日志信息，及时发现并处理潜在问题。

		行中的关键日志信息。	
多维度监控	验证平台的多维度监控功能，确保系统能够从多个维度展示资源和任务的使用情况。	客户维度监控：测试平台对不同客户的资源使用情况进行监控的能力，确保客户资源使用情况清晰可见。	系统能够从客户维度展示资源使用情况，用户能够清晰了解各个客户的资源使用状况。
		资源维度监控：测试对不同资源类型（如CPU、GPU、存储、网络等）使用情况的监控，确保资源使用情况透明、直观。	系统能够从资源维度展示各类资源的使用情况，用户能够清晰了解各类资源的消耗和利用情况。
		任务类型维度监控：测试对不同类型任务的资源消耗和执行情况进行监控的能力，确保系统能够全面反馈任务执行的效率和效果。	系统能够从任务类型维度展示任务的执行情况，用户能够分析不同类型任务的资源消耗和执行效率。

A.1.3 镜像管理

测试项目	测试目标	测试内容	预期结果
镜像管理功能测试	验证平台对镜像的管理功能，确保镜像上传、下载、版本管理和编辑操作的稳定性和有效性。	镜像上传：测试用户上传自定义镜像的功能，验证上传过程的稳定性及上传后镜像的可用性。	用户能够顺利上传自定义镜像，上传过程无中断，上传后镜像可正常使用。
		镜像下载：测试用户从仓库下载镜像的功能，确保镜像下载的稳定性。	用户能够成功下载所需镜像，下载后镜像可用。
		镜像版本管理：测试同一镜像的多版本管理功能，验证用户对不同版本镜像的切换、删除和回滚操作。	用户能够成功管理镜像的多个版本，并能切换、删除或回滚至指定版本。
		镜像属性编辑：测试镜像的属性编辑功能，验证镜像名称、描述、标签等属性的修改和保存操作。	用户能够成功编辑并保存镜像属性，修改后的属性在系统中正确显示。
镜像使用场景测试	验证平台在不同使用场景下对镜像的支持，确保镜像在任务运行中的稳定性和一致性。	工作空间启动：测试用户选择镜像作为工作空间启动镜像的功能，确保工作空间能够基于选定镜像成功创建。	用户能够顺利选择镜像并启动工作空间，工作空间环境与镜像配置一致。
		任务提交：测试用户在任务运行中选择镜像作为基础环境的功能，验证任务运行过程中镜像环境的稳定性。	用户能够顺利选择镜像并提交任务，任务运行过程中的环境与镜像配置一致，任务执行无异常。
		镜像一致性验证：测试镜像在不同任务和环境中的一致性，确保同一镜像在不同使用场景下表现一致。	同一镜像在不同任务和环境下表现一致，保证了环境的可复现性和任务的稳定性。

A.2 模型开发平台功能评估

A.2.1 模型训练与推理

测试项目	测试目标	测试内容	预期结果
模型训练支持	验证平台对不同规模和复杂度的模型训练任务的支持，确保训练任务的高效执行和稳定性。	单机单卡训练：测试在单机单卡环境下的模型训练，确保训练任务能够顺利执行并达到预期性能。	模型能够在单机单卡环境下顺利训练，训练过程稳定，性能达到预期。
		单机多卡训练：测试在单机多卡环境下的模型训练，验证多卡之间的数据同步和任务协调能力。	模型在单机多卡环境下训练时，多卡之间数据同步良好，任务协调顺利，训练效率提升显著。
		多机多卡分布式训练：测试在多机多卡分布式环境下的模型训练，确保任务在大规模环境中的执行效果。	模型在多机多卡分布式环境下训练时，各节点间任务协调良好，训练过程顺利且无明显瓶颈。

		预训练模型支持：测试平台对预训练模型的支持，验证常见预训练模型在平台中的加载和微调能力。	平台能够顺利加载和微调常见的预训练模型，微调过程高效且结果符合预期。
		训练监控：测试对训练过程的实时监控功能，确保用户能够实时查看训练进度、资源使用情况和性能指标。	用户能够实时监控训练过程中的进度、资源使用和性能指标，确保训练按计划进行。
模型推理支持	验证平台对模型推理任务的支持，确保推理任务的高效执行和部署稳定性。	单机推理：测试在单机环境下的模型推理任务，确保推理过程稳定且响应时间符合预期。	模型在单机环境下能够顺利执行推理任务，推理过程稳定且响应时间符合预期。
		分布式推理：测试在分布式环境下的模型推理任务，验证系统对大规模推理任务的支持和负载均衡能力。	模型在分布式环境下推理任务能够顺利执行，系统能够有效进行负载均衡，推理性能达到预期。
		推理框架支持：测试平台应支持 2 款以上推理框架支持，确保模型能够在不同推理框架下高效运行。	平台能够顺利支持多种推理框架，模型在不同框架下的推理性能良好，符合预期。
		推理服务部署与监控：测试推理服务的部署功能，确保推理服务能够快速扩展和缩减，实时监控推理服务的状态和性能指标。	推理服务能够顺利部署，并根据需求进行快速扩展和缩减，用户能够实时监控推理服务的状态和性能指标。
训练容错能力	验证平台训练任务容错能力支持，确保能够有效的开机自检、错误检查、任务恢复、日志输出。	开机自检：测试任务开启容错后是否开机自检，验证输出日志是否正确。	训练任务开启前支持训前检测，并输出检测日志。
		错误检测：测试任务异常后是否检查错误并定位，输出正确日志。	任务异常时正常进行错误检测，并且定位问题输出正确容错日志。
		任务恢复：测试任务异常后是否从上一个检查点恢复任务，确保任务能够按照配置的重启次数进行重启。	任务异常后从上一个检查点进行恢复，正常按照配置的重启次数进行重启。
		容错日志：测试容错日志是否正常输出，是否可读和正确。	容错日志正常显示开机自检和错误检测，并且保证可读性和正确性。

A.3 模型应用平台功能评估

A.3.1 预置大模型

测试项目	测试目标	测试内容	预期结果
预置大模型	验证平台预置的模型通过正确的配置可以被多种语言习惯的开发者直接调用，并获得模型返回内容	大语言模型 API 调用：基于平台提供的调用示例，配置有效的 API 密钥，输入请求内容	正常获得大模型的服务返回的请求信息、文本生成结果，和 token 统计
		大语言模型参数配置：基于有效的 API 密钥，配置不同的系统提示词、温度、最高 p 值、最高 k 值等参数，输入用户提示词内容	模型返回结果遵循系统提示词的内容，并根据不同的模型参数采样值，返回不同的生成内容
		文生图模型 API 调用：基于平台提供的调用示例，配置有效的 API 密钥，输入文本请求内容，获得生成队列信息，拉取生成结果	正常获得大模型的服务返回的图像生成任务信息和图像生成结果
		文生图模型 API 调用：基于平台提供的调用示例，配置有效的 API 密钥，输入图像信息，获得图像标识并提交生成任务，获得生成队列信息，拉取生成结果	正常获得大模型的服务返回的图像标识、图像生成任务信息和图像生成结果

		生图大模型参数配置：基于有效的 API 密钥，配置不同的正向、负向提示词，并配置种子值、步数、分辨率、调度等参数，发起生成需求	模型返回结果遵循正向、负向提示词的内容，并根据不同的模型参数采样值，返回不同的生成内容
		API 密钥身份验证：基于平台提供的调用示例，配置无效的 API 密钥，输入调用内容	服务接口返回身份验证不通过
		多编程语言 API 调用验证：基于平台提供的调用示例，配置环境，有效的 API 密钥，输入调用内容	多种编程语言的调用示例可正常获得大模型服务返回的结果
		OpenAI 兼容性：基于 OpenAI 已有接口的开发者，更换模型调用的地址、API 密钥和模型名称，其他正常输入调用	可正常基于 OpenAI 的调用方式返回结果
	验证平台提供多种能力使得用户可筛选定位到需要的模型	模型标签筛选：基于平台提供的模型类型、模型厂商、模型参数大小等标签，进行模型筛选	可筛选出所有符合筛选条件的模型
		模型信息查看：基于平台提供的模型卡片，查看模型简介信息、模型性能信息等信息	模型卡片的信息与模型本身信息相符无误。

A.3.2 模型体验

测试项目	测试目标	测试内容	预期结果
模型体验能力	验证平台提供直观的大语言模型验证能力，不需要通过脚本即可验证模型效果	大语言模型对话能力验证：基于平台提供的对话界面，对大语言模型提交输入指令内容	正常获得大模型的服务返回对话内容，可选择让模型重新生成结果，查看本轮对话使用的 token 数量
		大语言模型配置能力验证：基于基于平台提供的对话界面，调整平台暴露出来的可配置项，包括是否流式输出返回，最长的返回 token 数量，温度、最高 P 值等参数，提交输入指令	正常获得大模型的服务返回对话内容，返回内容的方式符合配置，返回的模型风格和多样性符合配置。流式输出过程中，可停止生成。
		大语言模型对比能力验证：基于平台提供的模型*芯片选择界面，配置两个以上模型与芯片组合，输入指令进行对话验证	选择的多个模型服务同时响应用户输入的同一个指令，开始返回内容。根据模型大小、使用芯片、模型配置等因素，多个模型与芯片的组合可能以不同的速度、不同的内容进行返回。
	验证平台提供直观的生图模型验证能力，不需要通过脚本即可验证模型效果	生图模型能力验证：基于平台提供的生图界面，选择随机的指令内容，提交生成	随机选择的指令内容包含正向和负向内容。 根据用户选择的输入指令内容正常返回生成的图片，可预览图片效果。
		生图模型验证：基于平台提供的生图界面，配置图片比例、图片数量、随机种子等参数，提交生成	根据用户配置的参数信息，同样的指令内容会返回不同的生图数量和生图效果。
		生图结果交互：基于平台生成的图片，进行复制、下载等进一步交互。	可复制到剪贴板。 可下载到本地。

A.3.3 模型微调

测试项目	测试目标	测试内容	预期结果

模型微调	验证平台提供大模型的微调能力，可基于用户数据集，发起自动化的微调任务，并对微调任务的生命周期做管理	大模型微调任务创建验证：基于平台提供的微调服务创建界面，完成所有必须内容的配置，提交创建微调任务	可完成任务的创建提交，可在微调服务任务列表，查看已创建的信息。
		大模型微调任务数据集格式可用性验证：在创建大模型微调任务时，可点击下载示例数据集。	自动发起微调示例数据集的下载。完成下载后可查看数据集提示词与响应 的文件内容。
		大模型微调任务配置验证：在创建大模型微调任务时，可选择使用的基础模型、微调方式、训练参数配置、数据集切分方式和使用的芯片类型，并提交创建微调任务	根据用户的不同配置，可成功发起不同的微调任务，并可通过微调服务任务列表，查看已创建的任务的信息字段。
		大模型微调任务生命周期查看验证：根据任务的训练阶段不同，展示任务的状态	用户可看到模型微调任务正常状态从创建中到最后运行完成的状态流程。
		大模型微调任务生命周期管理验证：根据任务的训练阶段不同，用户可对任务进行多种操作	用户可停止尚未完成的微调任务。 用户可编辑微调任务名称和描述。 用户可删除任务。

测试项目	测试目标	测试内容	预期结果
模型微调	验证平台提供大模型的微调能力，可查看用户已创建的任务信息和详细的训练信息。	大模型微调任务详情查看：用户点击任务名称，进入任务详情页，查看任务基本信息。	用户可在任务详情页获取微调任务提交时的配置信息，包括：基础模型、微调方式、参数配置以及数据集配置。 用户可查看训练过程中已经运行的时间、已经使用的 token 量、和剩余的时间
		大模型微调任务详情查看：用户点击任务名称，进入任务详情页，查看训练过程效果指标	在训练任务运行中，可查看每个步骤输出的损失值等效果指标
		模型微调任务详情查看：用户点击任务名称，进入任务详情页，查看训练事件。	在训练任务全生命周期，可查看不同的事件节点，包括创建过程、资源调度状态、数据处理状态、训练的每个周期完成节点，最终训练完成等信息。

A.3.4 非预置模型管理与部署

测试项目	测试目标	测试内容	预期结果
非预置模型管理与部署	验证平台提供大模型的模型部署能力，支持部署服务的创建与生命周期管理	大模型部署服务创建验证：基于平台提供的模型服务创建界面，完成所有必须内容的配置，提交创建模型服务	可完成模型服务的创建提交，可在模型服务列表，查看已创建的信息。
		大模型部署模型服务配置验证：在创建大模型模型服务时，可选择需要的模型的来源（微调任务）、查看使用的部署方式和资源来源。	根据用户的不同模型选择配置，可成功发起不同的微调任务的结果模型部署。同一个微调任务的模型可以被多次部署。
		大模型部署模型服务生命周期查看验证：根据服务部署进度，展示服务的状态	用户可看到模型服务正常状态从创建中到变为在线可用的状态流转。
		大模型部署模型服务生命周期管理操作验证：根据服务部署状态不同，用户可对服务进行多种操作	用户可停止部署中或在线状态的服务。 用户可启动已停止的服务。 用户可编辑模型部署的服务名称和描述。 用户可删除服务。

	验证平台提供大模型的部署模型服务能力，可查看用户已创建的模型服务和部署后的服务的可用性验证	大模型部署模型服务详情查看：用户点击模型服务详情，进入服务详情页，查看服务基本信息。	用户可在服务详情页获取部署模型时提交时的配置信息，包括：模型来源、资源来源、资源配置、实例数量 以及服务可被调用的在线地址
		大模型部署模型服务调用查看：用户点击服务名称，进入服务详情页，查看调用说明，进行调用验证	用户可查看调用说明，并使用调用示例验证部署服务的可用性 在线状态的服务会正常返回模型返回结果 非在线状态的服务无法正常返回模型结果
		大模型部署模型服务监控查看：用户点击服务名称，进入服务详情页，查看服务监控，查看监控指标	在服务有调用的情况下，用户可查看部署的服务所收到的 QPS、请求流量和响应情况、以及大语言模型场景的首字延迟、token 吞吐量、端到端延迟，token 总计等信息在无服务调用的情况下，无相关数据展示。
		大模型模型服务日志查看：用户点击任务名称，进入任务详情页，查看日志，进行查阅	用户可按时间翻阅服务部署的日志记录以及服务收到的调用日志记录。

A.3.5 用量统计

测试项目	测试目标	测试内容	预期结果
用量统计	验证平台提供完整的用量统计，能覆盖模型服务调用多时间段、多模型范围等多种维度的统计信息	新增模型调用统计验证： 通过 API 或体验中心使用模型，至少发生一次完整对话返回，点击进入用量统计界面	正常展示调用模型总数、调用 token 总数、调用服务总次数、异常调用总次数，其中除异常调用外，其他均大于等于 1
		无调用时统计验证： 全新账号无任何形式调用，点击进入用量统计界面	正常展示调用模型总数、调用 token 总数、调用服务总次数、异常调用总次数均为 0
		切换统计时间周期： 在有调用情况下，调整统计的时间周期，查看调用数据	选中统计周期内无调用时，调用次数统计值均显示为 0 选中统计周期内有调用时，调用次数统计值正常反应
		切换统计模型范围：在有调用情况下，调整不同的模型范围，查看调用数据	仅展示选中的模型范围的调用数据和调用信息
		异常调用统计： 在模型服务调用报错情况下，查看调用数据	异常调用总次数增加
	验证平台提供完整的用量统计，模型训练场景多时间段、多模型范围等多种维度的统计信息	新增模型训练统计验证： 通过微调服务支持完成一次模型微调任务，点击进入用量统计界面查看模型训练分页	正常展示已用基础模型总数、已训练 token 总数、模型训练总任务数、累计训练总时长，以上值均大于等于 1
		无完成微调任务统计验证： 无任何微调任务完成，点击进入用量统计界面	正常展示已用基础模型总数、已训练 Token 总数、模型训练总任务数、累计训练总时长，以上值均为 0
		切换统计时间周期： 在有微调任务已完成情况下，调整统计的时间周期，查看训练任务数据	选中统计周期内无微调任务完成时，训练相关统计值均显示为 0 选中统计周期内有微调任务完成时时，训练相关统计值正常反应
		切换统计模型范围：在有微调任务已完成情况下，调整不同的模型范围，查看训练任务数据	仅展示选中的模型范围的模型微调训练任务信息

A.4 性能评估

测试项目	测试目标	测试内容	预期结果

分布式训练性能	平台对比原有的裸机性能有没有折损	单机单卡性能测试：在单机单卡环境下，使用平台测试模型训练的 MFU 指标。	使用平台进行单机单卡的模型训练指标符合预期。
算力稳定性	验证大规模分布式模型训练可从中断恢复	1. 启动一个大规模分布式模型训练任务，使用至少4个节点，每个节点配备8个GPU。 2. 让训练任务运行2小时，记录训练进度和性能指标。 3. 模拟其中一个节点的硬件故障，如突然断电。 4. 观察平台的自动故障检测和重新调度行为。 5. 记录系统恢复时间和恢复后的训练进度。 6. 比较中断前后的训练性能和结果一致性。	系统应在 5 分钟内检测到故障并重新调度任务，恢复训练，且训练结果与无中断情况下的结果偏差不超过 1%。
平台稳定性	验证长时间运行下的资源管理效率	1. 在平台上同时启动多个不同类型的工作负载，包括训练任务、推理服务和开发容器，可以持续运行较长时间。 2. 每隔一段时间，增加或减少一个工作负载，模拟真实环境中的动态变化。 3. 持续监控平台的CPU使用率、内存占用、GPU利用率和网络吞吐量。 4. 记录开发框架的最小单元调度情况、资源分配和负载均衡效果。 5. 观察并记录任何系统报错、延迟增加或性能下降的情况。	系统应能稳定运行较长时间，资源利用率始终保持在70%-85%之间，无明显的性能衰减或资源泄露。

A.5 云平台安全评估

测试项目	测试目标	测试内容	预期结果
漏洞扫描	通过漏洞扫描软件确认平台不存在中高风险漏洞	使用 Acunetix Web Vulnerability Scanner（漏洞扫描工具）对平台进行漏洞扫描	扫描结果不存在中高风险漏洞
		使用 Xray（网络安全扫描工具）对平台进行漏洞扫描	扫描结果不存在中高风险漏洞
		使用 Goby（图形化漏洞扫描工具）对平台进行漏洞扫描	扫描结果不存在中高风险漏洞
		使用 Appscan（应用安全漏洞扫描工具）对平台进行漏洞扫描	扫描结果不存在中高风险漏洞
渗透测试	通过手工渗透测试确认平台不存在中高风险漏洞	通过模拟黑客真实攻击的方式测试平台可能存在的安全漏洞	渗透测试结果不存在中高风险漏洞
密钥管理	验证平台提供密钥管理能力，支持 API 密钥的创建、使用的生命周期	API 密钥的创建能力验证： 用户可点击创建 API 密钥，设置密钥的名称，提交创建。	可完成 API 密钥的创建，名称保持一致。 可创建多个 API 密钥。
		API 密钥的管理能力验证：针对已创建的 API 密钥，用户可点击启用、禁用 API 密钥	点击特定密钥的禁用按钮后,API 密钥状态变得已禁用。 点击特定密钥的启用按钮后,API 密钥的状态恢复为已启用。
		API 密钥的使用预先条件校验：新创建的 API 密钥，点击复制	未绑定手机号的用户在复制密钥时会弹窗要求先绑定手机号。 完成手机号绑定后会二次提交验证,需要用户输入手机验证码
		API 密钥的使用预先条件校验：复制 API 密钥时，输入需要的手机号验证码	输入正确的手机验证码可正常复制密钥 正常复制的密钥可被通过鉴权
		API 密钥的查看能力验证：用户进入密钥管理界面	可查看所有已创建的密钥和相关信息：名称、密钥头尾信息、创建时间、最近调用时间、状态

		API 密钥编辑能力：点击修改名称，用户可输入新的 API 密钥名称，点击提交	输入与其他密钥不同名的名称，提示修改成功，被修改的密钥名称发生即时变化 输入与其他密钥同名的名称时，提交后提示该名称已存在，修改不成功。

附录 B
（资料性）

智能算力云平台评估等级示例

智能算力云平台评估等级的基础级仅满足所有必选要求，提升级和引领级满足的可选要求数量可参考六西格玛管理统计学。其中，提升级满足所有必选要求且至少满足1个西格玛，26个可选要求的30.23%，即8条可选要求；引领级在提升级的基础上，还要至少1个西格玛，（26-8）x（1-30.23%），即21条可选要求。依据上述等级要求的具体评估等级示例见表B.1。

表B.1 智能算力云平台评估等级示例

能力域	能力子域	基础级	提升级	引领级
资源调度及管理功能	多租户管理	6.1	6.1	6.1
	算力管理及分配	6.2 a)～g)	6.2	6.2
	镜像管理	6.3 a)～c)	6.3	6.3
	平台监控	6.4 a)～c)	6.4	6.4
	基础资源调度	6.5 a)～c)	6.5 a)～c)	6.5
	分布式训练调度	6.6 a)～d)	6.6 a)～d)	6.6
	推理服务调度	6.7 a)	6.7 a)	6.7
	人工智能加速芯片复合调用	6.8	6.8	6.8
	弹性伸缩	6.9 a)，b)	6.9 a)，b)	6.9
	运营管理	-	-	6.10
模型开发功能	开发调试	7.1	7.1	7.1
	基础分布式任务	7.2	7.2	7.2
	模型训练与推理	7.3 a)，b)	7.3	7.3
模型应用功能	预置大模型	8.1 a)～d)	8.1	8.1
	模型体验	8.2 a)，b)	8.2 a)，b)	8.2
	模型微调	8.3 a)～c)	8.3 a)～c)	8.3
	非预置模型管理与部署	8.4 a)～f)	8.4 a)～f)	8.4
	用量统计	8.5 a)～c)	8.5	8.5
云平台性能	芯片算子优化性能	9.1 a)	9.1 a)	9.1 a)
	分布式训练性能	9.2	9.2	9.2
	调度性能	9.3	9.3	9.3
	算力稳定性	9.4.1 a)～c)	9.4.1 a)～c)	9.4.1 a)～c)
	平台稳定性	9.4.2 a)～d)	9.4.2 a)～d)	9.4.2 a)～d)
云平台安全	数据安全	10.1	10.1	10.1
	平台安全	10.2 a)～f)	10.2 a)～f)	10.2 a)～f)

附录 C
（资料性）

智能算力云平台等级自评报告模版

一、 被评估方基本信息

单位名称			
网址			
组织机构代码			
通信地址			
联系人姓名		手机/电话	
邮箱			

二、 评估团队信息

姓名	单位	部门	职务/职称	评估角色	联系方式

三、 评估依据

说明本次评估依据的国家法律法规及本标准条款。

四、 测试方法

说明本次评估的测试方法，包括测试环境、测试工具和测试流程等。

五、 测试结果

对本标准第 6～10 章的 88 条必选要求和 26 条可选要求逐条测试，分别记录测试结果（符合/不符合），并依据 5.4.3 节，向评估管理机构或第三方评估机构提供建议的审查方式（技术审查、实地检查、测试验证）。

评估维度	评估要求	可选/必选要求	符合/不符合	建议的审查方式
资源调度及管理功能评估	依据第 6 章条款，逐条评估			
模型开发功能评估	依据第 7 章条款，逐条评估			
模型应用功能评估	依据第 8 章条款，逐条评估			
云平台性能评估	依据第 9 章条款，逐条评估			
云平台安全评估	依据第 10 章条款，逐条评估			

六、 评估结论

依据评估结果，总结说明：

1） 是否符合所有的 88 条必选要求；

2） 符合的可选要求数量；

3） （可选说明）不符合相关要求的原因。

参 考 文 献

[1] GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求
[2] GB/T 31168-2023 信息安全技术云计算服务安全能力要求
[3] GB/T 36073-2018 数据管理能力成熟度评估模型
[4] GB/T 37988-2019 信息安全技术 数据安全能力成熟度模型
[5] T/SHSIC 0101-2023 智算中心算力性能评估测试方法
[6] T/SHSIC 0102-2024 智算中心验收能力规范